Méthodologie pour la constitution d'un corpus comparatif de narration orale en Occitan: objectifs, défis, solutions

Carruthers, J., & Vergez-Couret, M. (2018). Méthodologie pour la constitution d'un corpus comparatif de narration orale en Occitan: objectifs, défis, solutions. *Corpus*, 1-24. https://doi.org/10.4000/corpus.3490

Published in:
Corpus

Document Version:
Peer reviewed version

Méthodologie pour la constitution d'un corpus comparatif de narration orale en Occitan : objectifs, défis, solutions

Janice Carruthers, j.carruthers@qub.ac.uk

Marianne Vergez-Couret, m.vergez-couret@qub.ac.uk

Queen's University, Belfast

Dans cet article, nous souhaitons présenter et discuter de notre méthodologie pour la constitution d'un 'petit corpus' comparatif de narration orale en occitan (contes et récits), menée au sein du projet EXPRESSIONARRATION, subventionné par un *Marie Skłodowska-Curie Individual Fellowship* de deux ans (Horizon 2020).[1] Il s'agit d'un 'petit corpus' nouveau et unique, dans une langue minorisée, ce qui soulève un certain nombre de défis particuliers. Ayant expliqué et discuté des objectifs globaux du projet EXPRESSIONARRATION et des motivations pour la constitution de ce corpus en section 1, nous poserons en section 2 notre problématique en ce qui concerne la complexité des rapports entre l'écrit et l'oral dans la pratique du conte. En section 3, nous développerons plus en détail les défis méthodologiques posés par la constitution de ce type de petit corpus en langue minorisée et nos perspectives sur les 'solutions' que nous avons adoptées en ce qui concerne nos trois sous-corpus. Nous terminerons en section 4 avec un résumé de notre contribution au débat sur la problématique des petits corpus et une brève discussion des possibles utilisations de notre corpus pour des études sur l'occitan dans différents domaines (linguistique, littéraire et anthropologique) et sur les langues minorisées plus généralement, où la constitution de corpus numérisés pose souvent des difficultés.

1 Objectifs du projet : enjeux théoriques et pratiques

Le corpus est réalisé avec pour objectif de faciliter une exploration de la relation entre langage et oralité à travers une étude de la temporalité en occitan et en français. Il s'agit d'un corpus de narration orale, c'est-à-dire, de contes et récits qui représentent différents degrés d'oralité, ces degrés d'oralité étant étroitement liés à la complexité des rapports entre l'oral et l'écrit. Les théories linguistiques actuelles ont réussi depuis longtemps à dépasser la simple dichotomie oral/écrit (Koch et Oesterreicher, 2001 ; Crystal, 2001) et Carruthers (2005) a déjà abordé la question des différents degrés d'oralité pour le français à partir de l'analyse d'un corpus de contes traditionnels (contes transmis par voie orale dans une tradition orale) et un corpus de contes contemporains (contes performés oralement par des conteurs qui puisent leurs sources dans l'écrit), appelé le *French Oral Narrative Corpus*[2]. Les rapports oral/écrit sont particulièrement complexes, variables et fascinants dans le cas des contes et récits, en raison de plusieurs facteurs de variation comme la question des sources, le canal de transmission, la 'formation' (ou non) du conteur/de la conteuse, et le contexte de performance. Mais notre connaissance de cette complexité et de la façon dont les différents degrés d'oralité opèrent est encore inégale, et la description des traits des genres oraux encore sous-développée, voire même inexistante dans le cas de langues minorisées comme l'occitan. Pourtant, cette problématique est particulièrement pertinente dans le cas des contes en langues régionales, où la tradition orale est souvent moins ignorée dans la pratique des conteurs contemporains qu'elle ne l'est pour ceux qui content uniquement en français.

Pour constituer un corpus de contes et récits en occitan dans le but d'aborder la temporalité dans différents degrés d'oralité, nous avons exploité plusieurs sources de données, à savoir :

- des entreprises de sauvegarde du patrimoine qui ont débuté à la fin du XIX^{ème} et qui ont eu pour objectif de collecter des données orales et d'en publier des versions écrites ;
- des collectes qui ont été entreprises dans la deuxième moitié du XXème siècle dans un but de sauvegarde du patrimoine, de la littérature orale et de la langue occitane. Les interprètes

[1] EU projet 655034.

[2] www.frenchoralnarrative.qub.ac.uk

qui ont été enregistrés lors de ces collectes s'inscrivent dans une tradition où les contes ont été transmis de génération en génération de façon orale ;

- les performances de « nouveaux conteurs » qui puisent en partie leurs sources dans l'écrit (notamment grâce aux résultats de la première entreprise) ainsi que dans l'oral (notamment grâce aux résultats de la seconde entreprise). Ces conteurs et conteuses font partie d'un renouveau du conte mais ils se distinguent clairement des nouveaux conteurs de langue française qui puisent leurs récits exclusivement dans des sources écrites (Carruthers, 2005) ;
- pour les nouveaux conteurs et conteuses français(es), nous disposons du *French Oral Narrative Corpus*.

Dans le projet EXPRESSIONARRATION, nous souhaitons apporter une contribution originale au débat sur les modèles théoriques de l'écrit et de l'oral, en considérant en particulier les conditions de transmission et de performance qui viendront compléter les méta-caractéristiques traditionnellement prises en compte dans les modèles existants. Pour ce faire, l'accent sera mis sur une analyse sémantique et discursive des traits temporels, reconnus comme importants pour la structure linguistique des narrations orales, tels que les **temps verbaux**, les **indices de relations de discours** entre propositions narratives, et les **adverbiaux de localisation temporelle** pouvant jouer le rôle structurant de cadre de discours. Dans le cas des temps verbaux, nous analyserons en particulier le rôle discursif de **l'alternance des temps** dans les propositions narratives : par exemple, entre le prétérit et le présent narratif en occitan ou entre le passé composé et le présent narratif en français parlé. Les fonctions discursives de l'alternance temporelle dans les propositions narratives des narrations orales ont été analysées dans une grande variété de langues et peuvent être textuelles, métalinguistiques, structurales et expressives.[3] Dans le cas des **indices de relations de discours**, nous nous intéressons en particulier aux marqueurs linguistiques qui apparaissent fréquemment entre les propositions narratives, comme par exemple 'alavetz/alors' ou 'e puèi/et puis' qui peuvent fonctionner très différemment à l'écrit et à l'oral.[4] En ce qui concerne les '**cadres**', nous nous intéressons surtout aux adverbiaux temporels détachés en tête de phrase qui peuvent jouer un rôle structurel dans la mémorisation et la performance des contes, comme par exemple 'un jorn/un jour' ou 'lo lendoman matin/le lendemain matin'.[5]

Notre analyse de la temporalité emploiera plusieurs théories linguistiques, y compris la 'Segmented Discourse Representation Theory' (SDRT) et la théorie de l'encadrement.[6] Ces deux approches théoriques ont apporté des perspectives éclairantes à l'analyse de la temporalité dans le discours écrit dans une variété de langues. La SDRT aborde la cohérence sémantique des discours à travers une série de relations de discours (ci-après RD) telles que Narration, Elaboration, Arrière-plan... Elle a renoncé aux analyses qui attribuent une fonction particulière aux temps verbaux, connecteurs et adverbiaux (qui se sont souvent révélées inefficaces pour analyser les discours authentiques) pour adopter des analyses qui éclairent la façon dont les relations de discours sont exprimées linguistiquement. La SDRT offre également l'avantage de lier les analyses des phénomènes tels que les temps verbaux, les connecteurs et les adverbiaux, dont les fonctions dans l'établissement des RD sont généralement indissociables.[7] La théorie de l'encadrement concerne la description des cadres de discours qui vont fonctionner différemment dans des textes narratifs (où leur rôle est surtout structural) et non-narratifs (où leur rôle est plutôt temporel).[8] En revanche, le discours oral est rarement abordé à travers ces modèles et les langues minorisées ne figurent presque jamais dans la discussion. Nous faisons l'hypothèse que la complexité des différents degrés d'oralité attestés dans le cas des contes et récits, ainsi que les particularités linguistiques de l'occitan au niveau de la temporalité (p.e. le fait que le

[3] Voir par exemple Fleischman (1990), Bres (1998) et Carruthers (2005).

[4] P.e. Bras, Le Draoulec et Vieu (2001) et les articles dans un 'Special Issue' du *Journal of French Language Studies* (2005) intitulé: 'Discourse Organisation through time and space'.

[5] Voir Carruthers (2011) et les articles dans *Langue Française* (2005), numéro basé sur le thème : 'Les adverbiaux cadratifs'.

[6] Asher et Lascarides (2003) et Charolles (1997) respectivement.

[7] P.e. Bras, Le Draoulec et Vieu (2001) et Bras, Le Draoulec et Vieu (2003) sur le connecteur *puis*.

[8] Voir Charolles (1997) et Le Draoulec & Péry-Woodley (2005).

passé simple, longtemps abandonné dans le français parlé, est employé habituellement en occitan), créent un contexte particulièrement riche pour l'analyse des relations entre langage et oralité. Les objectifs globaux du projet peuvent se résumer donc comme suit : une analyse des phénomènes temporels dans un corpus de contes et récits qui contient différents types ou degrés d'oralité ; une comparaison avec un corpus de contes et récits français ; une exploration de la relation entre temporalité, mémorisation et performance ; une exploration du concept d'oralité dans ce contexte et une contribution innovatrice à l'étude des rapports entre langage et oralité.

Or, pour entreprendre ce projet, il faut disposer d'un corpus qui varie au niveau des sources, au niveau du canal de transmission et au niveau du contexte de performance. L'idéal serait de disposer d'un corpus numérisé en xml (pour garantir la compatibilité et la préservation à long terme), qui puisse recevoir des traitements linguistiques automatiques au moyen d'un analyseur morphosyntaxique et être annoté selon un des systèmes internationaux d'annotation comme la TEI (Text Encoding Initiative).[9] Aucun corpus existant ne correspond à cette vision. C'est donc avec ces objectifs en vue que nous développons notre méthodologie pour la constitution d'un petit corpus 'spécialisé' que nous appelons **OcOr** (Occitan, Oral), divisé en trois sous-corpus inédits, dont chaque sous-corpus connaît également une subdivision dialectale interne (cf. section 3.1) :

- **OWT** (Occitan, Written, Traditional) : versions publiées de contes et récits de tradition orale ;
- **OOT** (Occitan, Oral, Traditional) : versions collectées de contes et récits traditionnels transmis par voie orale dans une tradition orale ;
- **OOC** (Occitan, Oral, Contemporary) : performances enregistrées de contes et récits contemporains.

La comparaison français/occitan se fera sur la base d'une sélection de contes et récits du 'French Oral Narrative Corpus' (voir note 2) :

- **FOC** (Français, Oral, Contemporary) : performances enregistrées de contes et récits contemporains.

Dans cet article, nous nous focalisons sur le corpus OcOr (i.e. les trois sous-corpus OWT, OOT et OOC), et en particulier sur les défis de la constitution d'un petit corpus spécialisé en langue régionale. Mais nous aborderons également la comparaison des corpus OOC et FOC, surtout en ce qui concerne la variation dans les pratiques du conte. Les questions que nous nous posons dans l'article sont les suivantes :

- Quelles sont les complexités à prendre en considération dans l'analyse des relations entre l'oral et l'écrit dans le cas du conte en occitan ? En quoi les sources de conteurs, le canal de transmission et le contexte de la performance diffèrent d'un sous-corpus à l'autre ? Quel est l'impact de cette problématique sur nos décisions en ce qui concerne la construction de notre corpus ? Nous explorons ces questions en section 2.
- Quels sont les défis méthodologiques pratiques auxquels le chercheur doit faire face quand l'on veut construire un tel corpus, étant donné les multiples complications du contexte de cette langue minorisée ? Quelles solutions peut-on offrir à ces problèmes afin de construire le meilleur corpus possible, non seulement pour les besoins de notre projet mais aussi pour faciliter un emploi plus large du corpus par d'autres disciplines et pour aider la réflexion sur la construction d'autres petits corpus en langues minorisées ? Nous explorons ces questions en section 3.

La section 4 contient nos conclusions et nos réflexions sur les applications possibles de notre méthodologie.

2 Variations dans la pratique du conte : complexité théorique des relations entre l'oral et l'écrit

Nous présentons dans cette section les raisons pour lesquelles les contes et récits de nos quatre sous-corpus se définissent différemment en fonction du contexte dans lequel ils ont été transmis et diffusés. Ces différences sont évidemment fondamentales pour le raisonnement qui soutient la

[9] www.tei.org

constitution du corpus OcOr et il est donc crucial de les explorer en détail et de les prendre en considération, autant pour comprendre la complexité des relations entre l'oral et l'écrit dans le cas des contes, que pour pouvoir, par la suite, comparer certains traits temporels dans les différents sous-corpus. Il s'agit donc non seulement de constituer un petit corpus dans une langue minorisée mais également de créer trois sous-corpus, selon un certain nombre de paramètres que nous allons décrire dans cette section.

Les contes et récits constituant **le sous-corpus OOT** sont des exemples même de contes et récits de la tradition orale. Les enregistrements dont nous disposons sont le résultat de collectages par des enquêteurs auprès de conteurs et conteuses qui sont né(é)s entre 1860 et la première guerre mondiale.[10] Par définition, le conte traditionnel se caractérise d'une part par une « structure rigide », une « forme simple » qui s'élabore en respectant un ensemble de règles implicites et d'autre part par une « réalisation mouvante », une *performance*[11] chaque fois unique, même chez un seul conteur (Belmont, 1999, 10). L'exemple 1 illustre deux démarrages du même conte par le même conteur dans un même enregistrement, la reprise du conte étant provoquée par une panne de l'enregistreur :

1a) Alara un còp i aviá un òme, èra paure. E s'èra maridat e sabiá pas endont menar la femna, aviá pas cap d'ostau ni res. En se passegint, rencontrèt un òme bien vestit, galhard, que li diguèt :

(Alors une fois il y avait un homme, il était pauvre. Et il s'était marié et ne savait pas où emmener sa femme, il n'avait pas de maison ni rien. En se promenant, il rencontra un homme bien habillé, gaillard, qui lui dit :)

1b) Alara un còp i aviá un òme qu'èra paure. Aviá pas cap d'ostau ni res, e s'èra maridat e sabiá pas endont menar la femna. E trobèt un tipe que li diguèt :

(Alors une fois il y avait un homme qui était pauvre. Il n'avait pas de maison ni rien, et il s'était marié et ne savait pas où emmener sa femme. Il trouva un type qui lui dit :)

Le conte traditionnel n'a pas d'auteur particulier. Bru parle de contes et récits « ancestraux », ces derniers ayant été transmis de génération en génération sans que les conteurs n'en connaissent nécessairement l'origine (Bru, 2010, 33). Il s'agit d'une sorte d'œuvre collective, élaborée dans l'oralité, tout le temps renouvelée, ce qui explique qu'il existe une multitude de versions pour un même conte. Chaque performance, chaque version du conte est différente. Belmont parle du « caractère imparfait » du conte (Belmont, 1999, 11) dans la mesure où tout conteur peut omettre un épisode ou un motif ou emprunter des épisodes et des motifs de plusieurs contes pour en élaborer un. Le contenu mélange souvent des histoires venant du patrimoine européen ou même mondial avec des éléments locaux comme des noms de lieu ou de personnes. Comme l'exprime Marie-Louise Tenèze en parlant de son corpus de contes de l'Aubrac : 'venus des deux point opposés de l'horizon littéraire, de la réalité personnelle et de la fiction impersonnelle, les récits, dans le profil narratif concret d'une région, en arrivent à se côtoyer, voir à s'emprunter leurs fonctions. Ainsi s'explique comment ont pu m'être racontés, au chapitre des histoires vraies, tant de récits dont j'ai par la suite reconnu le caractère commun' (Tenèze 1975, 106).

Pour la constitution de notre **sous-corpus OWT**, nous nous intéressons également à la tradition orale, la grande différence avec OOT étant le fait qu'il s'agit ici de plusieurs entreprises d'édition de collectes qui ont eu pour objectif la mise en écrit et la publication de contes et récits. Il nous reste donc des versions écrites, publiées entre le milieu du XIXème siècle et le début du XX^{ème} siècle, créées à la base de collectages à une époque où les enregistrements ne sont pas possibles. Néanmoins, de notre point de vue, ces contes sont censés être 'proches' de l'oral, même si la mise en écrit entraîne nécessairement des modifications importantes du fait du changement de nature du canal, de l'oral vers l'écrit. Cette mise en écrit implique nécessairement une planification et doit répondre à des exigences sur les plans syntaxiques et discursifs auxquelles les productions orales ne sont pas soumises. Elle implique également le recours à divers procédés stylistiques afin de rendre compte des variations de voix, rythme, intensité, gestes et mimiques caractéristiques de l'oral. Chaque conte et récit publié va être ainsi inévitablement plus ou moins fortement marqué du style de l'auteur qui en a fait la mise en écrit. Pour la constitution de notre sous-corpus OWT, notre critère de sélection est la source du conte

[10] Pour un résumé des détails des trois sous-corpus, voir l'Annexe 1. Nous discuterons en plus de détail les sources de toutes nos données en Section 3.1.

[11] Calque du mot anglais *performance* pour désigner l'acte de narration de vive voix.

ou récit publié que nous souhaitons orale en premier lieu. Sont exclus les contes et récits littéraires, comme l'œuvre de Charles Perrault ou de Jean Boudou pour le domaine occitan, ainsi que les contes et récits que nous qualifions de 'création littéraire', autrement dit, qui n'ont pas été oraux en premier lieu, même si ces derniers ont été fortement inspirés de la tradition orale.[12] Néanmoins, la frontière entre ces deux grands sous-ensembles est loin d'être toujours nette. Etant donné que nous nous intéressons surtout à l'oralité, nous focalisons notre attention sur les versions où l'influence de la matière orale sur l'entreprise de mise en écrit du conte est la plus respectée. Pour résumer, nous avons donc veillé à sélectionner des contes et récits traditionnels, issus de collectes de contes de la tradition orale qui ont été transmis aux interprètes par voie orale, puis mis à l'écrit par les auteurs.[13]

Depuis les années 80, on assiste en France et dans le monde à un renouveau du conte, qui s'assimile à un mouvement artistique scénique.[14] Des conteurs amateurs ou professionnels montent sur scène, participent à des festivals, des concours, des rencontres. **Les corpus OOC et FOC** cherchent à rendre compte de la variété de ces nouvelles pratiques du conte en langue régionale et en langue française. Contrairement aux contes et récits du corpus OOT, transmis de façon spontanée lors du collectage, les conteurs contemporains préparent leur matériel (souvent en ayant recours à des sources écrites) et répètent leurs contes et récits en vue d'une performance publique. On bascule ici dans l'univers du spectacle avec, chez certains conteurs, le recours à une mise en scène (accessoires, instruments, décor…) plus ou moins élaborée. Tous ces « nouveaux conteurs » s'approprient le matériel, le retravaillent à leur manière avec une part de création et d'invention plus ou moins importante d'un conte à l'autre, ou d'un conteur à l'autre. Nous notons pourtant une différence dans les pratiques du conte en occitan et en français, notamment en ce qui concerne la profondeur de la relation des conteurs contemporains en langue occitane à la tradition orale. Contrairement aux conteurs uniquement de langue française dans le corpus FOC, les conteurs contemporains en langue occitane dans le corpus OOC accordent une place importante à la culture et à la langue régionale, puisant majoritairement leurs sources dans le répertoire traditionnel de leur région, autant dans les versions écrites qu'orales, selon ce qu'ils ont comme matériel à leur disposition (versions publiées de collectages, enregistrements, consultation des proches), même si, chez d'autres conteurs, dont il est fait allusion dans la citation ci-dessus, la part de création et d'invention peut prendre le dessus:

Alavetz un autre conte mes [rires] tradicional totjorn fin l'ai benlèu un pauc los cambi un pauc benlèu benlèu un pauc mes gaire mes soi pas capable d'inventar es pas donat a tot lo monde aquò si òm inventa totjorn un pauc e donc… (corpus ooc_vailet)

(Alors un autre conte mais [rires] traditionnel toujours enfin je l'ai peut-être un peu je les modifie un peu peut-être peut-être un peu mais peu mais je ne suis pas capable d'inventer ce n'est pas donné à tout le monde ça si on invente toujours un peu et donc…)

A la différence de OWT et OOT, la grande majorité des conteurs en occitan dans notre corpus OOC ne parlent pas l'occitan comme langue maternelle. Ils parlent tous français et occitan ; ils ont été scolarisés en français mais ils ont tous une expérience personnelle de l'apprentissage de l'occitan (voir la discussion dans 3.3).

Les contes et récits inclus dans FOC représentent l'activité contemporaine des conteurs dans un contexte authentique de performance et les transcriptions proviennent du *French Oral Narrative Corpus* (Carruthers, 2013). Tous les conteurs sont des « nouveaux conteurs » qui content en français, viennent de différentes régions de France et qui n'ont pas acquis leur répertoire dans la tradition orale, puisant leurs sources plutôt dans l'écrit à une échelle mondiale.[15]

Pour résumer, les trois sous-corpus représentent donc des degrés d'oralité différents, nous permettant, dans notre analyse des structures temporelles, d'explorer divers types de relation entre

[12] Belmont (1999 : 10) définit le conte littéraire comme une « œuvre littéraire […] régie par la volonté créatrice d'un individu », œuvre qui est plutôt définitive et close. Le conte littéraire a très largement été diffusé, en France, à travers notamment l'œuvre de Charles Perrault, jusqu'à presque en faire oublier l'essence même du conte traditionnel.

[13] Cette démarche de l'auteur est souvent décrite en préface des œuvres.

[14] Cf. Calame-Griaule (1991) et un ensemble de festivals organisés entre autres par le CLIO (Conservatoire Contemporain de Littérature Orale) et la FEST (Federation of European Storytelling).

[15] Voir Carruthers (2013) pour plus de détails sur la construction du *French Oral Narrative Corpus*.

l'oral et l'écrit. Tous les contes sont définis en tant que genre (structure relativement rigide régie par un ensemble de règles implicites (Belmont, 1999, 10)) mais seuls les contes et récits d'OOT ont été transmis dans la pure tradition orale. Les contes et récits d'OWT ont été collectés à partir de sources orales et produits en versions écrites publiées. Les performances d'OOC et FOC sont orales et spontanées mais les sources sont souvent écrites bien que cette influence soit moindre pour OOC étant donné la relation plus forte avec la tradition orale. Pour reprendre la terminologie de Zumthor (1983), OWT et OOT sont des exemples d'oralité 'mixte', dans la mesure où ils font partie d'une tradition orale dans une société qui a développé un système d'écriture mais où l'influence de celui-ci (pour des raisons sociologiques et éducationnelles) n'est que partielle. En revanche, OOC et FOC représentent plutôt une oralité 'seconde', fortement influencée par l'écrit, et par son prestige dans le cas du français. Ces facteurs de variation sont résumés dans le tableau 1 :

	OWT	**OOT**	**OOC**
Nature de la source	Orale	Orale	Orale ou écrite
Provenance de la source (niveau)	Locale	Locale	Régionale
Contexte de transmission (chez l'interprète)	Traditionnel	Traditionnel	Non traditionnel (source écrite et orale, création)
Préparation	Planification, exigences sur les plans discursifs et syntaxiques.	Spontanéité, peu de préparation : hésitations, répétitions, oublis.	Planification, préparation écrite et répétition orale. Spontanéité dans la performance.
Contexte de collectage	Collectage et mise en écrit	Collectage et enregistrement	Performance et enregistrement
Participants	Enquêté et écrivain	Enquêté et enquêteurs	Conteur amateur ou professionnel et public
Diffusion (nature du canal)	Ecrite	Orale	Orale
Structure rigide propre au conte/récit	✓	✓	✓
Réalisations plurielles	×	✓	✓
Attachement à la langue et à la culture régionale	✓	✓	✓

Tableau 1. Tableau comparatif des facteurs de variation dans OWT, OOT et OOC

Les facteurs de variation spécifiques à la comparaison des corpus de performances contemporaines en occitan et en français sont résumés dans le tableau 2 :

	OOC	**FOC**
Langues	Occitan	Français
Source (nature de la source)	Orale ou écrite	Ecrite
Provenance de la source (niveau)	Régionale	Mondiale
Attachement à la langue et à la culture régionale	✓	×

Tableau 2. Tableau comparatif des facteurs de variation dans OOC et FOC

3 Construire le corpus OcOr : défis méthodologiques

Pour la constitution de ce corpus, nous souhaitons profiter de toutes les avancées récentes en linguistique de corpus en ce qui concerne les outils de numérisation et d'annotation. Nous adoptons

donc les nouveaux standards en matière de diffusion en utilisant le format XML (TEI P5) qui garantit la pérennité et la réutilisabilité des données sur le plan international. Le corpus sera diffusé avec une Licence Creative Commons (BY NC SA) afin de garantir la meilleure diffusion possible pour les chercheurs de différentes disciplines. Néanmoins, constituer ce petit corpus spécialisé en langue régionale, avec un temps de constitution limité[16] soulève des défis particuliers qui sont peu discutés en linguistique de « grand corpus ». Dans cette section, nous discuterons de ces défis, tout en expliquant dans chaque cas, l'approche que nous avons adoptée. Les défis en question sont les suivants :

- Variation dialectale, cf. section 3.1 ;
- Absence totale de données textuelles numériques, cf. section 3.1 ;
- Variation diachronique, cf. section 3.3 ;
- Multitude des graphies employées dans les versions publiées, cf. sections 3.4 et 3.5;
- Adaptation d'outils d'annotation (Reconnaissance Optique de Caractère (OCR) et analyseur morphosyntaxique) à la spécificité des trois-sous corpus, cf. section 3.2 et section 3.6;
- La multitude des types de conte, cf. section 3.7.

3.1 L'occitan et la variation dialectale

Pour constituer un corpus en occitan, il faut se confronter à une variation interne relativement importante organisée en six dialectes principaux : auvergnat, gascon, languedocien, limousin, provençal et vivaro-alpin (Bec, 1995) sur un territoire couvrant le sud de la France, le val d'Aran en Espagne et douze vallées italiennes. L'occitan est une langue romane non unifiée et non standardisée dans son ensemble, et chaque dialecte connaît une variation interne propre.

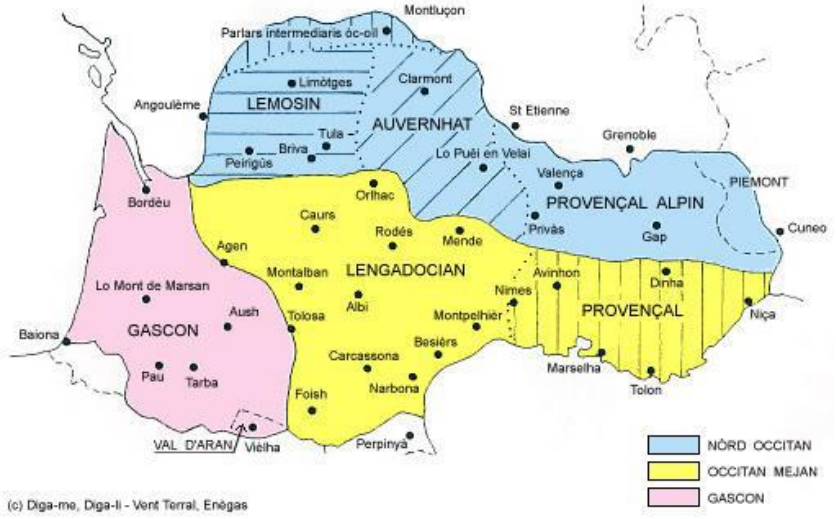


Figure 1. Aire linguistique occitane (Carles 2005, [11])

Nous avons pu réunir un corpus de contes et récits de deux grands ensembles dialectaux, le languedocien et le gascon, ce qui implique en quelque sorte l'imbrication de 'petits sous-corpus' dialectaux à l'intérieur des 'petits sous-corpus' OOC, OOT et OWT. En revanche, les trois sous-corpus ne sont pas nécessairement équilibrés du fait de la disponibilité et la qualité des sources ; il fallait trouver un bon équilibre entre ces facteurs. Le gascon est surreprésenté dans le corpus OWT qui contient 23 contes, écrits par 7 auteurs différents, dont 15 textes sont en gascon et 8 en languedocien. Par contre, le languedocien est surreprésenté dans le corpus OOT, où 19 enregistrements sont en languedocien et seulement 7 en gascon et dans le corpus OOC avec 8 enregistrements en languedocien et 5 en gascon. Etant donné notre choix du format XML et de la TEI, le corpus pourra à l'avenir être augmenté, d'une part pour améliorer l'équilibre entre les dialectes déjà présents et d'autre part pour étendre le corpus aux autres ensembles dialectaux pour lesquels il existe bien évidemment une tradition de littérature orale et un renouveau du conte avec des performances contemporaines.

[16] Neuf mois au début du Fellowship.

Pour démontrer les manifestations de la problématique de la variation dialectale et intradialectale présente dans nos sous-corpus, nous nous focalisons sur les traits temporels que nous souhaitons étudier. Par exemple, pour l'étude des adverbiaux de localisation temporelle et des indices temporels des relations de discours, nous aurons une importante variation lexicale, illustrée à travers quelques exemples, dans le tableau 3 :

Français	**Occitan**
un jour	un jorn, un dia
le lendemain	l'endoman, lo lendoman, lendoman, l'endeman, lo lendeman, lendeman, l'endedia
alors	alara, alavetz, aladonc, alòrs (forme française), lavetz
aussitôt	tanlèu, autanlèu

Tableau 3. Exemple de variation au niveau du lexique

La variation sera également très importante au niveau des formes verbales. Prenons dans le tableau 4, par exemple, le verbe français « dire » et quelques-unes des variations possibles dans notre corpus :

Lemmes	**Paradigmes au présent**	**Prétérit, 3ème personne du singulier**	**Information dialectale**
Dire	disi, dises, ditz, disèm, disètz, dison	Diguèt	Standard verb'òc[17] languedocien
Díser	disi, dises, ditz, disem, disetz, disen	Digó	Standard verb'òc gascon
Díser	disi, dises, ditz, disèm, disètz, dison	Digoc	Chez J.-F. Blader[18]
díser/ díder	dic/disi, dises, ditz, disem, disetz, disen	Dishot	Chez Arnaudin (cf. note 18)

Tableau 4. Variation morphologique dans le domaine verbal

Or, cette variation importante au niveau des formes entraîne à son tour des défis à relever pour les choix de transcription (cf. section 3.4 et 3.5) et pour la mise en œuvre de l'annotation automatique (cf. section 3.6) d'autant plus dans un contexte où les données textuelles numériques n'existent pas encore.

3.2 Absence totale de données textuelles numériques

L'on peut parler, dans le cas de chacun des trois sous-corpus de OWT, OOT et OOC, d'une absence totale de données textuelles numériques qui sont lisibles et manipulables par ordinateur. Pourtant, cette absence pose des problèmes différents dans chaque cas et demande donc des solutions diverses.

En ce qui concerne OWT, il n'existe, à notre connaissance, aucune version éditée électroniquement des ouvrages publiés de collectes de contes de tradition orale. Certains ouvrages, libres de droit, sont disponibles sous forme d'images, sur les sites de la Bibliothèque Nationale de France (la BNF) et du Centre Interrégional de Documentation Occitane (le CIRDOC). Deux grands projets, Gallica[19] et son équivalent occitan, Occitanica[20], ont mené, depuis quelques années, des campagnes de grande envergure de numérisation d'images. Les ouvrages ainsi mis à disposition sur les sites web Gallica et Occitanica sont au format pdf et constitués d'une image par page d'excellente qualité. Cette image est généralement accompagnée de la sortie d'un logiciel de reconnaissance optique de caractères (OCR, *Optical Character Recognition*) qui, bien que non disponible à la visualisation, permet néanmoins à l'utilisateur de faire des recherches dans le texte/image. La figure 1 est une image extraite de *Contes*

[17] Le Verb'Òc est un conjugateur occitan mis à disposition par *Lo congrès permanent de la lenga occitana*, http://www.locongres.org/fr/applications/2014-02-03-17-16-34-fr/verboc-recherche

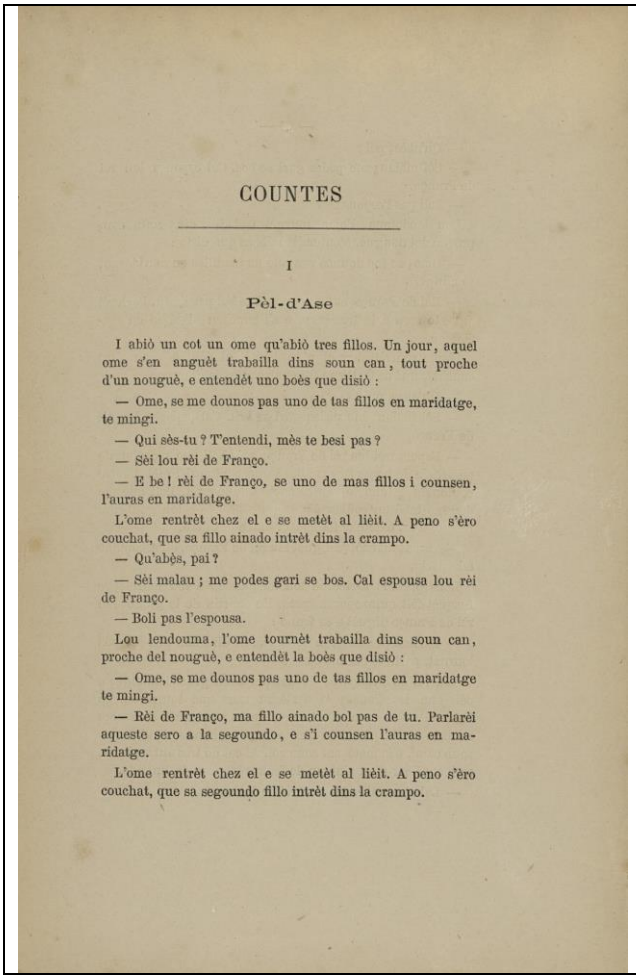
[18] Conjugaison de référence pour certains auteurs de notre corpus listée dans Grosclaude et Gilabèrt, N. (éds.), 77-80.

[19] http://gallica.bnf.fr/

[20] www.occitanica.eu

populaires recueillis en Agenais par J.-F. Bladé, disponible sur le site Occitanica, accompagnée de la sortie OCR (logiciel commercial ABBYY FineReader) qui nous a été fournie par le CIRDOC. Pourtant, la sortie OCR, quoique de très bonne qualité, n'est pas une réplique fidèle du texte de l'image. Les zones identifiées sur fond gris dans la figure 2 sont des erreurs de reconnaissance de l'image. Pour créer une version numérique de très bonne qualité, il faut donc, au préalable, passer par certaines étapes importantes : p.e. corriger les erreurs OCR, rétablir la ponctuation, corriger les erreurs dans la mise en forme (cf. les tirets des dialogues), reconstituer les mots tronqués etc.

Or, la difficulté à surmonter pour l'océrisation d'une langue comme l'occitan est l'absence d'outils libres de droit adaptés à la langue dans sa variété dialectale et graphique. Dans le cadre du projet Occitanica, le CIRDOC sous-traite une société de service qui emploie le logiciel commercial ABBYY FineReader. Employant une approche différente, Vergez-Couret *et al.* (2017) présentent divers types d'outils et diverses stratégies pour l'océrisation des langues « peu dotées » comme l'alsacien et l'occitan sans recourir à l'achat d'un outil commercial qui peut être en même temps coûteux et opaque quant à son mode de fonctionnement dans le cas des dialectes et graphies non-standards. Il est possible d'utiliser soit des outils développés pour des langues proches graphiquement (i.e. qui disposent du même alphabet) tels que Tesseract (logiciel libre développé par Google); soit des outils génériques probabilistes qui utilisent des techniques d'apprentissage automatique supervisés comme Jochre (développé par Joliciel). Un tel outil a été développé dans le cadre du projet ANR RESTAURE (RESsources informatiques pour le Traitement AUtomatique des langues REgionales) pour l'occitan écrit en graphie classique et en graphies non standards (Urieli et Vergez-Couret, 2013 ; Vergez-Couret *et al.*, 2017) et évalué avec des performances similaires à ABBYY pour la constitution de notre corpus OWT (Vergez-Couret, 2017). Les sorties des deux logiciels ABBYY et Jochre ont donc été exploitées pour constituer ce sous-corpus. Le second défi à relever vis-à-vis de la numérisation de OWT concerne alors l'hétérogénéité des graphies employées, défi discuté dans la section 3.4 sur la variation graphique.



COUNTES

I

Pèl-d'Ase

I abiò un cot un ome qu'abiò tres fillos. Un jour, aquel ome s'en anguèt trabailla dins soun can, tout proche d'un nouguè, e entendèt uno boès que disiò :

— Ome, se me dounos pas uno de tas fillos en maridatge, te mingi.

— Qui sès-tu ? T'entendi, mès te besi pas ?

— Sèi lou rèi de Franço.

— E be ! rèi de Franço, se uno de mas fillos i counsen, l'auras en maridatge.

L'ome rentrèt chez el e se metèt al lièit. A peno s'èro couchat, que sa fillo ainado intrèt dins la crampo.

— Qu'abès, pai ?

— Sèi malau ; me podes gari se bos. Cal espousa lou rèi de Franço.

— Boli pas l'espousa.

Lou lendouma, l'ome tournèt trabailla dins soun can, proche del nouguè, e entendèt la boès que disiò :

— Ome, se me dounos pas uno de tas fillos en maridatge te mingi.

— Rèi de Franço, ma fillo ainado bol pas de tu. Parlarèi aqueste sero a la segoundo, e s'i counsen l'auras en ma-ridatge.

L'ome rentrèt chez el e se metèt al lièit. A peno s'èro couchat, que sa segoundo fillo intrèt dins la crampo.

GOUNTES

I

Pèl-d'Ase

I abiò un cot un ome qu'abiò tres fillos. Un jour, aquel ome s'en anguèt trabailla dins soun can, tout proche d'un nouguè, e entendèt uno boès que disiò :

Ome, se me dounos pas uno de tas fillos en maridatge, te mingi.

Qui sès-tu ? T'entendi, mès te besi pas ?

Sèi lou rèi de Franço.

E be! rèi de Franço, se uno de mas fillos i counsen, l'auras en maridatge.

L'ome rentrèt chez el e se metèt al lièit. A peno s'èro couchat, que sa fillo ainado intrèt dins la crampo.

Qu'abès, pai?

Sèi malau; me podes gari se bos. Cal espousa lou rèi de Franço.

Boli pas l'espousa.

Lou lendouma, l'ome tournèt trabailla dins soun can, proche del nouguè, e entendèt la boès que disiò :

Ome, se me dounos pas uno de tas fillos en maridatge te mingi.

Rèi de Franço, ma fillo ainado bol pas de tu. Parlarèi aqueste sero a la segoundo, e s'i counsen l'auras en ma-ridatge.

L'ome rentrèt chez el e se metèt al lièit. A peno s'èro couchat, que sa segoundo fillo intrèt dins la crampo.

Figure 2. Une page extraite de *Contes populaires recueillis en Agenais* par J.-F. Bladé – image et sortie OCR

Le sous-corpus **OOT** est réalisé à partir de collectes qui ont eu lieu dans les années 1960-2000 et qui sont hébergées, sous forme d'enregistrements, au COMDT (Conservatoire occitan des musiques et danses traditionnelles) à Toulouse. Nous avons veillé à sélectionner uniquement des collectages contés (et non lus ou récités) ; tous les enregistrements sont des collectages effectués généralement sur le lieu de résidence de l'interprète. Bien que les enregistrements soient disponibles pour l'ensemble des collectes, ils ont rarement fait l'objet de transcriptions et lorsque des transcriptions ont été réalisées, elles ne répondent pas à nos exigences de fidélité à la source orale. Il est donc nécessaire de trier, dans une masse très importante, et de choisir le matériel ayant la meilleure qualité technique possible et étant le mieux renseigné du point de vue des métadonnées, avant d'entreprendre le travail de transcription (voir 3.4 et 3.5).

Pour créer le corpus **OOC,** les enregistrements de nouveaux conteurs en performance sont rares et pas toujours libres de droit. Quelques enregistrements sont disponibles au COMDT, auxquels nous avons ajouté un DVD.[21] Pour compléter ce corpus, il a donc fallu organiser des soirées de contes. Ceci a été réalisé à l'occasion de deux spectacles à Toulouse les 28 septembre et 26 octobre 2016, organisés en partenariat avec la Tuta d'Òc[22] (librairie occitane de Toulouse) dans le cadre du Festival Occitania organisé par l'Institut des Etudes Occitanes (IEO) et majoritairement financés[23] par le projet EXPRESSIONARRATION. Nous avons veillé à ce que la sélection des contes de ce corpus respecte la variété des nouvelles pratiques du conte en langue régionale, étant donné la complexité de la relation des conteurs contemporains à la tradition orale, discutée en section 2 : le mélange de contes traditionnels et créations contemporaines, une mise en scène plus ou moins élaborée, les sources (écrite ou orale) de transmission du conte… Tous ces enregistrements ont dû donc été transcrits par la suite (voir 3.4 et 3.5).

Les paramètres de variations autour de la problématique oral/écrit sous-tendant la constitution des trois sous-corpus a inévitablement une implication sur d'une part les périodes représentées dans chaque sous-corpus et d'autre part les générations de locuteurs impliqués, ce qui pose un défi particulier autour de la gestion de la variation diachronique et de la qualité des métadonnées sur les locuteurs.

3.3 Variation diachronique

Les contes et récits de nos trois sous-corpus ont été produits à des périodes différentes et auprès de personnes appartenant à trois générations de locuteurs différentes. La qualité des métadonnées pertinentes est variable mais l'on peut néanmoins généraliser en disant qu'il s'agit de locuteurs nés au XIXème siècle pour les contes publiés du corpus OWT,[24] d'interprètes nés entre le milieu du XIX^{ème} siècle et la Première Guerre Mondiale pour OOT,[25] et de locuteurs nés après la Seconde Guerre Mondiale pour OOC.[26] Or, ces différences de génération influencent profondément non seulement les pratiques du conte (voir la discussion dans section 2), mais aussi les compétences linguistiques en occitan des locuteurs, qui varient d'une génération à l'autre. Pour généraliser, on peut dire que les locuteurs de OWT et de OOT sont des locuteurs natifs de l'occitan qui ont des capacités en français très variables, voire inexistantes dans plusieurs cas, surtout dans OWT. En revanche, les locuteurs de OOC, nés après la 2nde guerre mondiale, parlent français et occitan mais les modes de transmission de l'occitan (oral/écrit, passif/actif) sont plus variés. Dans certains cas, l'occitan est transmis par les parents ou les grands-parents mais dans d'autres cas, l'occitan est appris de façon plus tardive comme

[21] Voir l'Annexe pour les détails.
[22] http://www.latutadoc.com/
[23] Nous remercions l'Institut des Etudes Occitanes (IEO), organisateur du Festival Occitania, qui a contribué généreusement à l'organisation et à la communication des deux événements.
[24] C'est une généralisation établie à partir des métadonnées fournies sur 4 locuteurs nés entre 1810 et 1824 et des années de publication des œuvres entre le milieu du XIXème et le début du XXème siècle.
[25] C'est également une généralisation établie à partir des métadonnées fournies sur 6 locuteurs nés entre 1876 et 1908 et des années de collectage depuis le milieu du XX^{ème} siècle jusqu'au début du XXI^{ème} siècle.
[26] Voir l'Annexe 1.

langue seconde.[27] Etant donné les deux niveaux de variation diatopique (cf. section 3.1) et diachronique, combinés à une petite taille de corpus, l'étude des phénomènes linguistiques, ici temporels, doit se faire lexème par lexème (même en cas d'apparente synonymie comme pour la paire *alara (languedocien)/alavetz (languedocien/gascon)* qui se traduit *alors* en français), zone géographique par zone géographique, voire individu par individu.

3.4 Variations graphiques

Le défi de la variation graphique concerne uniquement le corpus écrit OWT, constitué à partir d'ouvrages publiés en occitan (ou en bilingue français/occitan) entre le milieu du XIXème et le début du XXème siècle, époque où il n'y a plus de consensus sur une graphie unifiée de l'occitan. C'est le recul des publications à la fin du Moyen-Âge qui entraîne la perte des usages orthographiques médiévaux alors stables. Au milieu du XIX^{ème} siècle, Frédéric Mistral codifie pour le dialecte provençal une orthographe unifiée dite mistralienne mais cette époque est surtout caractérisée par une multitude de graphies non standards individuelles qui néanmoins partagent la caractéristique d'être basée sur les relations phonie/graphie de l'orthographe française.

Pour unifier d'un point de vue graphique l'ensemble des contes et récits issus de ces ouvrages, et pour faciliter l'étape d'analyse morphosyntaxique automatique (voir 3.6), nous avons opté pour une nouvelle convention graphique appelée graphie classique, élaborée depuis le milieu du XXème siècle avec pour inspiration la graphie des troubadours. Cette graphie a été élaborée avec pour objectif l'atténuation des différences dialectales tout en respectant les particularités de chaque dialecte (Sibille, 2007). Par exemple un seul graphème 'j' sera utilisé bien qu'il donne lieu à des prononciations différentes selon les dialectes [ʒ], [d͡z], [d͡ʒ], [j]. Pour le corpus OWT, le passage de la graphie originale (non standard) à la graphie classique a été fait manuellement (voir 3.5). Ensuite, pour faciliter la comparaison entre les différents sous-corpus de notre corpus, nous avons logiquement opté pour cette graphie pour l'ensemble de notre corpus OcOr, i.e. pour la transcription des corpus oraux OOT et OOC.

3.5 Choix de transcription

Pour le corpus écrit OWT, l'étape de la transcription consiste à passer de la graphie individuelle de l'auteur à une version en graphie classique, afin non seulement de créer une harmonisation graphique à travers les différents textes mais aussi pour faciliter la mise en œuvre de l'étape d'analyse morphosyntaxique automatique, pour laquelle des outils et des ressources ont déjà été développés pour l'occitan en graphie classique.

La graphie non standard individuelle de l'auteur se base sur les rapports graphie/phonie du français. La graphie, ainsi que des connaissances sur les caractéristiques phonétiques du parler, permettent généralement d'avoir une idée fiable sur la prononciation. Par exemple, « Qu'ére praoube » dans le parler landais d'Arnaudin devrait se prononcer [kerəpraubø] (avec des variations possibles sur le r, uvulaire [ʁ] ou apical [r]). Mais la question est plutôt de savoir quelle forme classique nous retenons. Dans ce cas, nous pouvons retenir la forme « Qu'èra praube ». Ici la graphie classique oppose les caractéristiques les plus communes du gascon et les particularités régulières du parler en question, notamment au niveau des voyelles dans « qu'**è**r**a** praub**e** » :

Voyelle graphique	Majoritairement prononcée	Prononcée dans ce parler
è	/ɛ/	/e/
a	/ɔ/	/ə/
e	/e/	/ø/

Tableau 5. Prononciation de trois voyelles

[27] On pourrait peut-être parler de 'néo-locuteurs' dans certains cas, dans la mesure où ces locuteurs ont appris la langue comme langue 'seconde'. Mais le concept de 'néo-locuteur' est problématique dans ce cas, étant donné la relation forte pour certains locuteurs avec la continuité d'une tradition linguistique et la possibilité que certains locuteurs soient locuteurs passifs avant d'être locuteurs actifs. Pour un projet contemporain sur les 'new speakers', voir http://www.cost.eu/COST_Actions/isch/IS1306.

Le choix opéré ici consiste à atténuer la différence dialectale, étant donné que la même forme graphique « Qu'èra praube » sera prononcée différemment d'un parler à l'autre. Ce choix peut aussi concerner les consonnes, comme par exemple pour les formes en graphie non standard « ouarda », « ouayre » et « ouaita ». L'initiale, ici prononcée avec la diphtongue /wɑ/, est majoritairement prononcée dans l'ensemble de la Gascogne avec la consonne suivie de la diphtongue /g wɑ / : nous avons alors opté pour les formes normalisées « guardar », « guaire », « guaitar ».

En revanche, certaines variations ne sont pas réductibles à la graphie, comme la forme en graphie non standard « neugue », que nous avons transcrit « negue », au plus proche de la prononciation dans ce parler, même si la forme la plus courante en occitan est « negre ». En règle générale, nous nous appuyons sur l'ensemble des variantes dialectales qui ont été répertoriées dans les dictionnaires en graphie classique. Les gallicismes sont conservés, comme par exemple le mot « même », mais transcrits en suivant les règles orthographiques de la graphie classique, ici « mèma ». Quand il demeure des doutes sur la façon de transcrire une forme originale, nous le signalons avec l'appareillage XML TEIP5 à notre disposition.

Pour les corpus oraux OOT et OOC, la transcription utilise donc la graphie classique et cherche à rester fidèle, autant que possible, suivant la clarté des enregistrements, aux choix phonétiques des interprètes, sur le même principe que décrit ci-dessus, soit en atténuant la différence dialectale quand elle est régulière, soit en respectant la particularité du dialecte quand elle est anecdotique. Les gallicismes sont également conservés avec donc une adaptation aux règles orthographiques de la graphie classique, par exemple les formes « formilhs » et « formilhèira » sont transcrites telles quelles, même si ce ne sont pas des formes standardisées (qui serait respectivement « formiga », « formiguièr »).

3.6 Adaptation des outils de traitement automatique des langues

Pour la constitution de ce corpus, nous avons souhaité utiliser deux types d'outils de traitement automatique des langues (TAL) : l'OCR dont nous avons parlé dans la section 3.2 et un outil d'analyse morphosyntaxique automatique dont nous parlons dans cette section. Cette étape consiste à attribuer automatiquement à chaque mot du corpus une étiquette morphosyntaxique: le lemme, la partie du discours (nom, adjectif, adverbe, verbe…) et des informations morphosyntaxiques associées (genre, nombre, mode, temps…). Cette annotation automatique est largement employée pour les grands corpus mais déployer ces outils dans le cas des langues minorisées nécessite quelques ajustements. Pour notre corpus, c'est une étape importante qui est un préalable à l'annotation des temps verbaux.

Un exemple d'annotation morphosyntaxique est présenté dans le tableau 6 sur la phrase « Un jorn, qu'arriba davant ua pòrta » (Un jour, il arrive devant une porte) :

Forme	Lemme	Catégorie	Informations morphosyntaxiques
un	un	Déterminant	masculin, singulier, indéfini
Jorn	jorn	Nom	commun, masculin, singulier
,	,	Ponctuation	
qu'	que	Particule énonciative	
Arriba	arribar	Verbe	présent de l'indicatif, 3ème personne du singulier
Davant	davant	Préposition	
Ua	un	Déterminant	féminin, singulier, indéfini
Pòrta	pòrta	Nom	commun, féminin, singulier
.	.	Ponctuation	

Tableau 6. Exemple d'annotations morphosyntaxiques sur une phrase

A l'heure actuelle, pour constituer des analyseurs morphosyntaxiques, les méthodes par apprentissage supervisé sont fortement plébiscitées. Ces méthodes visent à dégager des règles générales à partir d'exemples particuliers et nécessitent alors des données annotées pour l'entraînement et optionnellement un lexique. En ce qui concerne l'occitan en graphie classique, de nombreuses

ressources (corpus annotés et lexiques) ont été constituées dans le cadre des projets BaTelÒc (Bras et Vergez-Couret, 2016) et RESTAURE (Vergez-Couret et Urieli, 2015, Bernhard *et al.*, 2018). Dans un premier temps, ces ressources ont été constituées pour entraîner des modèles pour l'occitan languedocien avec le logiciel Talismane (Urieli et Tanguy, 2013). De nouvelles ressources ont ensuite été constituées, en suivant la logique suivante : étendre la création de ressources à d'autres variantes languedociennes, avant d'étendre la création de ressources à un autre dialecte, le gascon (Vergez-Couret, 2017). Nous disposons de peu d'éléments sur la façon d'adapter les outils d'annotation disponibles aux langues connaissant une grande variation interne. Une stratégie possible consiste à traiter chaque dialecte ou chaque parler comme une langue à part entière et de constituer pour ce dialecte des ressources spécifiques de très grande qualité (corpus annotés, lexiques ou grammaires selon l'approche choisie), ce qui nécessite des moyens humains et financiers très importants. Pour notre projet, le corpus OcOr contient des textes de deux grands ensembles dialectaux, à savoir, le languedocien et le gascon. Nous présentons dans Vergez-Couret (2017) des expériences visant à mettre au jour la meilleure stratégie pour l'annotation morphosyntaxique de ces deux dialectes avec Talismane et les ressources disponibles à l'heure actuelle, ce qui nous a mené à exploiter les similarités entre les dialectes et a employé un corpus multi-dialectal (languedocien et gascon) pour l'entraînement de Talismane dans le but de faire une pré-annotation de notre corpus.

Utiliser un analyseur morphosyntaxique pour faire une pré-annotation des temps verbaux dans nos corpus est un gain de temps considérable. Premièrement, l'analyseur morphosyntaxique propose une désambiguïsation des formes comme « pòrta » qui peuvent être soit un nom, soit un verbe selon les contextes. Deuxièmement, l'analyseur morphosyntaxique s'appuie sur une liste de formes répertoriées comme verbes conjugués dans un lexique, mais peut également repérer automatiquement, en se basant sur des probabilités de distribution des étiquettes de catégorie, des formes verbales non répertoriées dans le lexique. Cela est effectivement très fréquent avec toute la variation dialectale possible en occitan et particulièrement dans le domaine verbal, comme par exemple dans « quan estot mòrta » (quand elle fut morte), « estot » est une variante de « estó ». Seule la deuxième forme est répertoriée dans le lexique mais Talismane est capable de signaler « estot » comme un verbe dans ce contexte.

3.7 Variétés des types de conte

Il existe une variété importante de types de conte, la classification la plus importante pour le conte européen étant celle d'Aarne et Thompson (1961). Cette classification, employée régulièrement dans les domaines d'anthropologie et de littérature orale, a pour objectif de regrouper plusieurs variantes d'un même conte, autant dans des catégories larges (comme 'contes d'animaux', 'contes facétieux', 'contes religieux') que dans des sous-catégories (par exemple, les contes religieux sont subdivisés en 'dieu récompense et punit', 'la vérité vient au jour', 'l'homme dans le ciel' et 'l'homme promis au diable') et dans des catégories plus fines où il est possible d'attribuer un numéro précis au conte-type. En ce qui concerne notre corpus, nos priorités sont (i) la constitution d'un corpus contenant une variété de types de conte, même s'il est impossible d'avoir exactement la même représentation typologique dans chaque sous-corpus ; (ii) la possibilité de donner un maximum de métadonnées typologiques pour faciliter l'emploi du corpus par d'autres disciplines. Il était donc intéressant pour notre corpus de classer et de varier les types de conte, au moins dans les catégories larges définies par Aarne et Thompson. Chaque type est régi par ses propres règles et ses propres structures narratives (Propp, 1973) : par exemple, les contes formulaires sont plus contraints avec la reprise de phrases ayant la même structure syntaxique du début à la fin du conte tandis que les contes merveilleux ont tendance à être plus longs. Nous avons pu varier ces divers types de conte dans notre corpus, tout en sachant qu'il sera impossible de constituer un corpus homogène et représentatif, d'une part en raison de la petite taille du corpus et également étant donné que la quantité disponible de contes pour chaque type peut varier fortement d'un contexte à l'autre. Nous avons donc enrichi le corpus avec des informations dans les métadonnées sur le ou les conte(s)-type(s) pour les contes traditionnels des corpus OOT et OWT, de même que pour les contes puisés dans le répertoire des

contes traditionnels pour le corpus OOC, même si ces derniers n'ont pas été transmis dans un contexte de pure oralité, de sorte que l'on puisse identifier des variantes d'un sous-corpus à l'autre.[28]

4 Conclusions et applications

Cet article a exposé les motivations et les complexités de la constitution du corpus OcOr, petit corpus en langue minorisée qui contient trois sous-corpus, chacun apportant ses propres défis à plusieurs niveaux. Les complexités en question concernent la problématique théorique des rapports oral-écrit, les difficultés méthodologiques rencontrées dans le cas d'une langue minorisée avec la sélection de certains sous-ensembles dialectaux (transcription, graphie, application des outils de traitement de texte), le manque de données accessibles et de métadonnées dans certains cas ainsi qu'une absence quasi totale de données textuelles numériques. Un tel contexte de variation diatopique et diachronique pose également des défis de comparabilité et de représentativité et l'étude des phénomènes étudiés devra être menée prudemment zone géographique par zone géographique, voire individu par individu. Nous espérons donc avoir contribué au débat sur les défis de la constitution des petits corpus en langue minorisée et avoir créé, de plus, un ʿpetit corpusʾ spécialisé, le premier de son type, qui servira de base de données pour notre projet EXPRESSIONARRATION sur la relation entre le langage et l'oralité, à travers une exploration de la temporalité.

Pourtant, la constitution de ce petit corpus a des objectifs plus ambitieux au-delà de l'application immédiate à notre projet et c'est pour cette raison qu'il a été crucial d'utiliser un système de numérisation XML (la Text Encoding Initiative, TEIP5) qui est largement déployé à l'échelle internationale ainsi que la technologie la plus contemporaine en ce qui concerne l'harmonisation graphique, l'annotation morphosyntaxique et la capture des métadonnées. Un de nos objectifs plus globaux est que ce corpus soit une source riche de données pour la recherche sur la structure linguistique de l'occitan plus généralement. Il serait parfaitement réalisable, par exemple, d'ajouter d'autres couches d'annotation pour une variété de phénomènes syntaxiques tels que la négation ou l'ordre des mots, ou pour des phénomènes lexicaux. Ces phénomènes pourraient, grâce à ce corpus, être analysés non seulement dans des textes de différents degrés d'oralité mais aussi dans des textes qui représentent une certaine évolution diachronique, de la fin du 19ème siècle jusqu'au début du 21ème siècle. Nous espérons également que nos stratégies en ce qui concerne les difficultés particulières dans le cas des langues minorisées (transcription et numérisation des données face à la variation dialectale et la variation graphique, adaptation des outils de traitement automatique) serviront de modèle méthodologique pour d'autres projets sur des langues minorisées où les données disponibles ne sont pas adaptées à la méthodologie de la numérisation contemporaine. Enfin, nous espérons que ce corpus de contes et récits, émanant de différents contextes d'oralité, servira de ressource précieuse pour la recherche dans d'autres disciplines, notamment dans le domaine des études littéraires, et surtout en littérature orale et en anthropologie.

5 Références

Aarne, A., Thompson, S. (1961). *The Types of the Folktale, A Classification and Bibliography* (Helsinki : Academia Scientiarum Fennica).

Asher, N., Lascarides, A. (2003). *Logics of Conversation* (Cambridge : CUP).

Bec, P. (1995). *La langue occitane* (Paris : Que sais-je).

Belmont, N. (1999). *Poétique du conte. Essai sur le conte de tradition orale* (Paris : NRF Gallimard).

Bernhard, D., Ligozat, A.-L., Martin, F., Bras, M., Magistry, P., Vergez-Couret, M., Steiblé, L., Erhart, P., Hathout, N., Huck, D., Rey, C., Rosset, S., Sibille, J. (à paraître). 'Corpora with Part-of-

[28] Les détails des types de contes dans les quatre sous-corpus se trouvent dans l'Annexe 1.

Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard', dans *Proceedings of Language and Resources and Evaluation Conference LREC 2018*, Miyazaki, Japan.

Bras, M., Le Draoulec, A., Vieu, L. (2001). 'French adverbial 'puis' between temporal structure and discourse structure', dans Bras, M., Vieu, L. (éds), *Semantic and Pragmatic Issues in Dialogue : Experimenting with Current Theories* (Amsterdam : Elsevier), 109-46.

Bras, M., Le Draoulec, A., Vieu, L. (2003). 'Connecteurs et temps verbaux dans l'interprétation temporelle du discours : le cas de puis en interaction avec l'imparfait et le passé simple', dans Mellet, S., Vuillaume, M. (éds), *Modes de repérages temporels* (Amsterdam : Rodopi), 71–97.

Bras, M., Vergez-Couret, M. (2016). 'BaTelÒc: A text base for the Occitan language', dans Ferreira, V., Bouda, P. (éds), *Language Documentation and Conservation in Europe* (Honolulu: University of Hawaï Press), 133-149.

Bres, J. (1998). 'De l'alternance passé composé/présent en récit oral conversationnel', dans Borillo, A., Vetters, C., Vuillaume, M. (éds), *Variations sur la référence verbale* (Amsterdam : Rodopi), 125-136.

Bru J. (2010). 'De l'oral à l'écrit : la rupture', *Port Acadie : revue interdisciplinaire en études acadiennes / Port Acadie: An Interdisciplinary Review in Acadian Studies*, 16-17 : 33-43.

Calame-Griaule, G. (1991). *Le renouveau du conte* (Paris : CNRS Editions).

Carles, S. (1986). *Diga-me, diga-li : metòde audiòvisual d'ensenhament de l'occitan pels dròlles : lengadocian* (Enèrgas : Vent terral).

Carruthers, J. (2005). *Oral Narration in Modern French. A Linguistic Analysis of Temporal Patterns* (Oxford : Legenda).

Carruthers, J. (2011). 'Temporal framing in the *conte*. From theoretical debate to oral story performance', *French Studies*, 65 : 488-504.

Carruthers, J. (2013). *The French Oral Narrative Corpus* (www.frenchoralnarrative.qub.ac.uk).

Charolles, M. (1997). 'L'encadrement du discours: univers, champs, domaines et espaces', *Cahiers de recherche linguistique*, 6 : 1-73.

Crystal, D. (2001). *Language and the Internet* (Cambridge : CUP).

Fleischman, S. (1990). *Tense and Narrativity. From Medieval Performance to Modern Fiction* (London-New York : Routledge).

Grosclaude, M., Gilabèrt, N. (1998). *Répertoire des conjugaisons occitanes de Gascogne* (Orthez : Per Noste).

Koch, P., Oesterreicher, W. (2001). 'Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit', dans Holtus, G., Metzeltin, M., Schmitt, C. (éds), *Lexicon der Romanistischen Linguistik I.2* (Amsterdam-Atlanta : De Gruyter), 584–627.

Le Draoulec A., Pery-Woodley M.-P. (2005). 'Encadrement temporel et relations de discours', *Langue française*, 148 : 45-60.

Propp, V. (1970). *Morphologie du conte* (Paris : Seuil).

Sibille, J. (2007). 'L'occitan, qu'es aquò ?', *Langues et Cité : bulletin de l'observation des pratiques linguistiques,* 10 : 2.

Tenèze, M.-L. (1975). *Aubrac, V : Littérature orale narrative* (Paris : CNRS).

Urieli, A., Tanguy, L. (2013). 'L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane', dans *Actes de la 20ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013),* Les Sables d'Olonne, publication en ligne : https://halshs.archives-ouvertes.fr/halshs-00953754.

Vergez-Couret, M. (2017). 'Constitution et annotation d'un corpus écrit de contes et récits en occitan', *Analyses et méthodes formelles pour les humanités numériques,* ISTE OpenScience, 1-1, publication en ligne : https://www.openscience.fr/Constitution-et-annotation-d-un-corpus-ecrit-de-contes-et-recits-en-occitan.

Vergez-Couret, M., Bernhard, D., Urieli, A., Bras, M.; Erhart, P., Huck D. (2017). 'Numérisation et océrisation de textes pour les langues régionales : regards croisés sur l'occitan et l'alsacien', dans Chevry Pébayle, E. (éd), *Systèmes d'organisation des connaissances et humanités numériques* (London : ISTE Editions), 250-269.

Vergez-Couret, M., Urieli, A. (2015). 'Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan', dans *Actes du 2ème Workshop sur le Traitement Automatique des Langues Régionales de France et d'Europe (TALaRE'2015),* Caen, publication en ligne : http://talnarchives.atala.org/ateliers/2015/TALaRE/talare-2015-long-007.pdf.

Vergez-Couret, M., Urieli, A. (2014). 'POS-tagging different varieties of Occitan with single-dialect ressources', dans Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J. (éds), *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects* (Dublin : Association for Computational Linguistics and Dublin City University), 21-29.

Zumthor, P. (1983). *Introduction à la poésie orale* (Paris: Seuil).

Annexe 1

Le Corpus OcOr (OWT, OOT, OOC) et le Corpus FOC

	OWT	OOT	OOC	FOC
Nombre de contes/récits	23	26	13	13
Types de conte avec nombre	8 contes merveilleux, 3 contes facétieux, 3 contes formulaires, 2 récits légendaires, 3 contes d'animaux, 2 contes de l'ogre	7 récits légendaires, 5 contes merveilleux, 7 contes et récits facétieux, 4 contes d'animaux, 1 conte formulaire, 1 conte nouvelle	2 contes merveilleux, 1 conte de l'ogre dupé, 2 contes d'animaux, 2 récits avec êtres fantastiques, 1 récit légendaire, 1 conte facétieux, 1 récit-cadre, 3 créations contemporaines.	4 contes et récits facétieux, 3 récits légendaires; 2 contes merveilleux, 2 contes d'animaux, 1 conte de l'ogre dupé et 1 conte nouvelle.

	dupé 1 conte religieux, et 1 conte nouvelle.	et 1 conte de l'ogre dupé.		
Longueur (minutes ou nombre de mots)	18200 mots.	110 minutes : entre 28 secondes et 12 minutes, moyenne de 4 minutes. 16600 mots.	130 minutes : entre 3 et 20 minutes, moyenne de 10 minutes. 19000 mots.	120 minutes : entre 3 minutes et 18 minutes, moyenne de 9 minutes. 20000 mots.
Variétés dialectales	15 Gascon, 8 Languedocien	7 Gascon, 19 Languedocien	5 Gascon, 8 Languedocien	français
Source des données/état des métadonnées	BNF (Gallica), CIRDOC (Occitanica). Métadonnées : date de publication ; rien ou très peu sur les conteurs (parfois nom et/ou lieu).	COMDT, avec permission. Métadonnées lacunaires. Date de collectage, parfois, lieu ; parfois nom ou date de naissance.	Un enregistrement est extrait du DVD *Les contes du placard* de Florant Mercadier et exploité avec sa permission. Tous les autres enregistrements ont été effectués lors de deux événements organisés, les 28 septembre 2016 et 26 octobre 2016 à l'Ostal. Metadonnées détaillées.	www.frenchoralnarrative.qub.ac.uk Métadonnées détaillées.
Sexe des auteurs/conteurs et conteuses	Lacunaires (dont 6 H et 5 F)	12 F 6 M	2F 3M	7F 3M
Période de naissance des conteurs/es	XIX^{ème} siècle.	Entre 1860 et la première guerre mondiale.	Après la deuxième guerre mondiale.	Après la deuxième guerre mondiale.
Période de collectage/performance/ /publication	Milieu du XIXème jusqu'au début du XXème siècle.	1960-2000	2013-	1990 -